Queue Inferencing in M/M/1 Queues with Cumulative Departure Information

by

Vishal Sharma

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KANPUR

May 1997

Queue Inferencing in M/M/1 Queues with Cumulative Departure Information

A Thesis Submitted

in Partial Fulfillment of the Requirements

for the Degree of

Master of Technology

by

Vishal Sharma

to the

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

May 1997

dedicated to

my loving sisters

Certificate

This is to certify that the work contained in the thesis entitled **Queue Inferencing in M/M/1 Queues with Cumulative Departure Information** by **Vishal Sharma** has been carried out under my supervision and that this work has not been submitted elsewhere for a degree

Dr D Manjunath
Assistant Professor
Deptt of Electrical Engineering
Indian Institute of Technology
Kanpur

Acknowledgments

A lot goes into the making of a thesis by way of encouragement support inspiration and enlightenment I shall take this opportunity to echo my gratitude for all those who helped me in this effort

My first and foremost thanks are due to my thesis supervisor Dr D Manjunath for his indispensable guidance motivation and encouragement and his instant' one liners which helped me keep my spirits high to no end

Words fail to express my gratitude towards my parents and my sisters who so lovingly remain an unflinching source of encouragement in everything I do

During my stay in IIT/K I had the good luck to enjoy the company of a lot of good friends and well wishers Avi, Ashish Abhijit Alkesh Manoj ji, M P Agarwal (MPA), Mohit, Naresh, Pratima Santosh Shishir Sunita Subir, Vipin, Viveks (Awasthi Sharma Ranjan), Tomar and all other friends especially in the Telecom group, made my stay here a memorable one

Thanks also to Sandhya ji for her help whenever I required it

I would have loved to say thanks to everyone individually who helped and encouraged me in one way or the other, but one has to put a full stop somewhere My thanks to all such whom I failed to squeeze into the above list

Vishal Sharma

Abstract

Packet delays and queue lengths are important indicators of the quality of service offered at a node in a telecommunications network These parameters cannot be easily measured This necessitates the estimation of these parameters from other more easily available information like the traffic arrival rates to nodes the departure instants of packets or the cumulative departure counts Queue inferencing is a technique to estimate the queueing parameters using this kind of transactional data

Many queue inferencing algorithms have been developed All of these algorithms require detailed transactional data of the service initiation and termination instants of packets In this thesis we have proposed queue inferencing schemes that require less detailed data to estimate the packet delays and queue lengths We concentrate on the queue inferencing using the cumulative departure count information This information is more easily available from the network management information bases and is also less informative This information is collected by the network manager by polling the node at regular intervals

In the first method that we study we divide the polling interval into cycles comprising of an idle period and its adjacent busy period We then distribute the total departures among these busy periods to generate the kind of data required by existing queue inferencing algorithms We then use this data in the queue inferencing algorithms and estimate the waiting times We study the performance of this technique by evaluating the errors and the bias in the estimates by comparing the estimates with the 'real values from simulation

Next we derive an $O(d)$ formula for estimating the queue length at the end of a given time interval, which we call as the "residual' queue length, using the information of the cumulative departure count, d We derive an expression for the joint probability distribution of the cumulative departures and the residual queue length, and use this result to derive our estimation algorithm We also present some numerical results for our algorithm

Contents

List of Tables

Lıst of Figures

Chapter 1

Introduction

To guarantee a minimum quality of service for the applications running over a network we need proper management Among the most important parameters defining the quality of service are the delays that the data packets experience en route from source to destination and the queue lengths encountered by the packets at network nodes A direct measurement of these parameters is difficult We can, however, estimate these parameters from more easily available data like the data of packet departure instants at a node or the departure counts, or similar data about the traffic transacted at the node Queue inferencing is the technique for estimation of queue parameters using such transactional data

Various network management applications, like SNMP agents provide different types of data about the nodes they are monitoring *e g* the number of data packets transacted in a given interval and the number of erroneous data packets received We may also log more detailed transactional data like the arrival and departure instants of the packets processed at a node The cost of gathering detailed data is a higher consumption of network resources and the cost paid towards its dissemination to the managing agent is even higher in terms of these resources Hence algorithms requiring a less detailed set of data assume importance The earlier queue inferencing algorithms require detailed transactional data of the departure instants and busy period initiation instants and hence has limited practical utility In this thesis, we have developed methods to estimate queue parameters from the information about the cumulative departures in a time interval This is more easily obtained from the SNMP agents and also consumes less network bandwidth during dissemination

1 1 Motivation

Queue inferencing techniques estimate queueing information using the transactional data of a queueing system e g a node in a data network The incoming packets at the node are the customers which join the queue and after being processed leave the system A common model for the network traffic incident at a node is a Poisson process and that for the node is an M/M/1 queueing system [KLECN] Various performance parameters of the node may be estimated using queue inferencing algorithms

A major drawback of the existing queue inferencing algorithms is the requirement of detailed transactional data which limits the practical utility of these algorithms This has motivated us to develop some queue inferencing techniques requiring less detailed transactional data

In this thesis we first present a scheme to estimate the waiting times of customers in a queue using known queue inferencing algorithms We use the easily obtainable transactional data of the cumulative departure count during a time interval as the input in the scheme This data is less informative and our scheme is therefore less accurate than the case when the detailed service initiation and termination information is available

Next we develop an $O(d)$ algorithm for estimating the queue length at the end of a given time interval called the ‘residual queue length We derive an expression for the joint distribution of residual queue length and the cumulative departure count in a given time interval, and use it to derive our algorithm This algorithm assumes the knowledge of the utilisation of the server

1 2 Organisation of the Thesis

In this thesis, we propose queue inferencing schemes to estimate the waiting times of customers in an M/M/1 queue and discuss their error performance We later develop an algorithm for estimating residual queue lengths conditioned on the cumulative departure information We discuss these separately

In chapter 2 we discuss the previous work on queue inferencing techniques We then propose our scheme for estimation of waiting times of customers We discuss the errors in the estimate by comparing with simulation results We also propose some correction techniques to this scheme In chapter 3 we derive an $O(d)$ formula to estimate the queue lengths at the end of a given time interval using the cumulative departure count information After discussing the approach we derive the joint distribution of the residual queue length and the cumulative departure count We then develop our estimation algorithm Our algorithm assumes the knowledge of arrival rate of customers to the queue We discuss the estimation technique to measure the utilisation of the server Finally chapter 4 summarizes the work and identifies avenues for future work

Chapter 2

Queue Inferencing . Theory and Algorithms

2 1 Introduction

Consider a queueing system for which the transactional data are recorded in the form of service commencement and service completion times for each customer served This transactional data when rank ordered, allow the identification of busy and idle periods in the queueing system Larson [Lar90] proposed a method to estimate the transient queue lengths during a busy period from this transactional data Since the publication of this result algorithms to infer queue lengths and waiting times from transactional data have been proposed In the next section we discuss the early work of Larson [Lar90] and the algorithms proposed by Bertsimas and Servi [BeS92] In section 2 3 we discuss the algorithms developed by Daley and Servi [DaS92] using Markov chain techniques and taboo probabilities In section 2 4 we discuss the feasible arrival vector model of Manjunath and Molle [MaM96] Finally in section 2 5 we present a scheme to use queue inferencing techniques with only cumulative departure count information We develop the theory and present some experimental results for this scheme

2.2 Early Work

Larson [Lar90] proposed two algorithms to estimate the queue lengths during a busy period of an FCFS queue. He based his algorithms on the following two observations.

1. The ending of each busy period can be identified by a service completion time which is not immediately followed by a new service commencement. The subsequent service initiation is the beginning of a new busy period.

2. In a busy period, a service commencement at time t_i implies that the arrival time of the corresponding customer must be between the arrival time of the previous customer and t_i. Further, if the arrival process is known to be Poisson, the *a posteriori* probability distribution of the arrival time of this customer must be uniformly distributed in the specified interval.

The approach focuses on a single busy period. Since the completion (or commencement) of a busy period constitutes a renewal point in any Poisson arrival queue, the solution for one busy period is also the solution for any time period having an arbitrary number of busy periods. Larson used these observations to derive $O(n^5)$ and $O(2^n)$ algorithms to compute the transient queue lengths during an n customer busy period.

Bertsimas and Servi [BeS92] proposed an $O(n^3)$ algorithm which estimates the queue lengths during a busy period. A brief discussion of their work follows.

Consider an n customer busy period of an FCFS single server queue with Poisson arrivals that started at time t_0. Let τ_i be the (unknown) arrival time and t_i the service completion time of the ith customer. Let $N(t)$ be the cumulative number of arrivals in time $[t_0, t)$ and $Q(t)$ be the number of customers in the queue at time t^-. Without loss of generality we can define $t_0 \triangleq 0$. In a busy period, a service commencement at time t_j implies that the $(j+1)$th customer must have arrived before the departure of the jth customer. Let $O(\underline{t})$ be the event $\{0 < \tau_2 < t_1, \tau_2 < \tau_3 < t_2, \ldots, \tau_{n-1} < \tau_n < t_{n-1}\}$. Let $O(\underline{t}, n) \triangleq O(\underline{t}) \cap \{N(t_n) = n\}$.

If the arrival process to the queue is Poisson, the *a posteriori* probability distribution of the jth customer must be uniform in the interval between the arrival instant of the $(j-1)$th customer and its departure instant. Thus, the joint density of the event

$\{\tau_2 = x_1 \quad \tau_n = x_{n-1}\}$ conditional on $N(x_n) = n$ is given by

$$f(\tau_1 = x_1 \qquad \tau_n = x_{n-1} | N(x_n) = n) = \frac{(n-1)!}{x_n^{n-1}} \tag{2 1}$$

Therefore

$$Pr\{\tau_2 \leq t_1, \qquad \tau_n \leq t_{n-1} | N(t_n) = n\} = \frac{(n-1)!}{t_n^{n-1}} \int_{x=0}^{t_1} \int_{x_3=x_2}^{t_2} \qquad \int_{x=x_{n-1}}^{t} dx_2 \qquad dx_n \tag{2 2}$$

Then the estimate of the cumulative number in system at time t $0 \leq t \leq t_n$ for an n customer busy period can be shown to be given by

$$E[N(t)|O(\underline{t}\ n)] \;=\; (1-\theta)E[N(t_{j-1})] + \theta E[N(t_j)] \quad \text{for } t_{j-1} < t \leq t_j \tag{2 3}$$

where

$$\theta = \frac{t - t_{j-1}}{t_j - t_{j-1}} \tag{2 4}$$

Here $E[N(t_j)]$ is the expected number of cumulative arrivals at the departure instant of the jth customer in the busy period $E[N(t_j)]$ is given by

$$E[N(t_j)] \;=\; \sum_{k=j}^{n} kPr\{N(t_j) = k|O(\underline{t}\ n)\} \tag{2 5}$$

Efficient methods have been derived for the calculation of $Pr\{N(t_j = k|O(\underline{t}\ n)\}$ by Bertsimas and Servi [BeS92] The queue length estimate at time t $Q(t)$, is given by

$$Q(t) \;=\; N(t) - j \quad \text{for } t_{j-1} < t \leq t_j \tag{2 6}$$

Bertsimas and Servi also generalized the algorithm for the case of stationary interarrival times from an arbitrary distribution They also proposed an $O(n)$ on line algorithm to dynamically update the current estimates for queue lengths after each departure

2 3 Taboo Probabilities Daley and Servi

Daley and Servi [DaS92] exploited Markov Chain techniques and used taboo probabilities to estimate queue lengths from transactional data They constructed an exact $O(n^3)$ and an approximate $O(n^2 \log n)$ algorithm for busy periods with n customers for a single server FCFS queue They used the observation that the queue is never empty at any service completion instant (except that of the last customer) in a busy period In other words, the queue being empty at any departure instant inside a busy period is a taboo event. The queue length at the instant of service completion of the rth customer inside an n customer busy period ($r < n$) may then be expressed in terms of Markov chain transitions comprising only of non taboo events since the beginning of the busy period. They considered an n customer busy period of a single server FCFS queue with service times $S_1, \ldots, S_n$. S_i's are known and no assumptions are made about their statistics. Let t_0 denote the beginning of the busy period and for $r = 1, \ldots, n$ define

$$t_r = t_0 + S_1 + \ldots + S_r = t_{r-1} + S_r \tag{2.7}$$

Observe that t_r (t_{r-1}) is an epoch of service completion (commencement) of the rth customer in a busy period, the values of $S_1, \ldots, S_n$ are known (hence the t_r's are known). t_n is the epoch of busy period completion.

Let $N(t)$ be the number of customers in the queue at time t (excluding the customer in service) at time t, $t_0 \leq t \leq t_n$. $N(t)$ is left continuous, i.e. $N(t) = N(t^-)$. $N(t_0) = 1$ and $N(t_n) = 0$, and the event $N(t_s) = 0$ for $t_0 \leq t_s < t_n$ is a taboo event. The arrival process to the queue is Poisson. (We do not need to know the rate of the arrival process.) Define a subset of the complement of the taboo event as

$$A^{r_1, r_2} \triangleq \{N(t_s) > 0,\ s = r_1, \ldots, r_2\} \tag{2.8}$$

A^{r_1, r_2} is the event that the busy period did not end between t_{r_1} and t_{r_2}. By the left continuity of $N(t)$, $N(t_r) \triangleq N_r = N(t_r^-)$, and $N(t_r^+) = N(t_r) - 1$. Then, for $j = 1, 2, \ldots$ and $r = 1, \ldots, n-1$,

$$\begin{aligned} p^r_{j|n} &\triangleq Pr\{N_r = j | N_0 = 1, N_s \geq 1 (s = 1, \ldots, n-1), N_n = 0\} \\ &= Pr\{N_r = j | N_0 = 1, A^{1, n-1}, N_n = 0\}\} \\ &= \frac{Pr\{N_r = j, A^{1, n-1}, N_n = 0 | N_0 = 1\}}{Pr\{A^{1, n-1}, N_n = 0 | N_0 = 1\}} \end{aligned} \tag{2.9}$$

$p^r_{j|n}$ is the probability of N_r being equal to j in an n customer busy period ($r < n$), excluding all taboo events.

Using Feller's [FELL] notation for taboo probabilities, we have

$$\begin{aligned} {}_0p^{0,r}_{1,j} &= Pr\{N_s > 0 (s = 1, \ldots, r-1), N_r = j | N_0 = 1\} \\ &= Pr\{A^{1, r-1}, N_r = j | N_0 = 1\} \end{aligned} \tag{2.10}$$

${}_0p^{r_1\,r_2}_{j_1\,j_2}$ is the probability of the queue being in state j_1 at the r_1th departure instant and in state j_2 at the r_2th departure instant ($r_2 > r_1$) with the queue never being empty at any intermediate departure instant. Then

$$\begin{aligned} {}_0p^{r\,n}_{j\,0} &= Pr\{N_s > 0 (s = r+1, \ldots, n-1), N_n = 0 | N_r = j\} \\ &= Pr\{A^{r+1,n-1}, N_n = 0 | N_r = j\} \end{aligned} \tag{2.11}$$

Using the Markovian property of $\{N_r\}$ the numerator of (2.9) can be expressed as

$$Pr\{N_r = j, A^{1,n-1}, N_n = 0 | N_0 = 1\} = {}_0p^{0\,r}_{1\,j}\;{}_0p^{r\,n}_{j\,0} \tag{2.12}$$

Using the Chapman Kolmogorov equations, the denominator of (2.9) can be expressed as

$$\begin{aligned} Pr\{A^{1,n-1}, N_n = 0 | N_0 = 1\} &= {}_0p^{0\,n}_{1\,0} \\ &= \sum_{h\geq 1} {}_0p^{0\,r}_{1\,h}\;{}_0p^{r\,n}_{h\,0} \end{aligned} \tag{2.13}$$

Therefore, we have

$$p^{r}_{j|n} = \frac{{}_0p^{0\,r}_{1\,j}\;{}_0p^{r\,n}_{j\,0}}{{}_0p^{0\,n}_{1\,0}} \tag{2.14}$$

As $N_r \geq N_{r-1} - 1$ and $N_n = 0$ for a busy period of length n,

$$p^{r}_{j|n} = {}_0p^{r\,n}_{j\,0} = 0 \qquad \text{for } j > n - r \tag{2.15}$$

Chapman Kolmogorov equations are used to compute the taboo probabilities recursively as given below for $r = 1, \ldots, n$ and $j = 1, 2, \ldots$

$$\begin{aligned} {}_0p^{0\,r}_{1\,j} &= Pr\{A^{1,r-1}, N_r = j | N_0 = 1\} \\ &= \sum_{l=1}^{j+1} Pr\{A^{1,r-2}, N_{r-1} = l | N_0 = 1\}\, Pr\{N_r = j | N_{r-1} = l\} \\ &= \sum_{l=1}^{j+1} {}_0p^{0\,r-1}_{1\,l}\, Pr\{Y_r = j - l + 1\} \end{aligned} \tag{2.16}$$

while

$${}_0p^{0\,0}_{1\,j} = \delta_{1j} = \begin{cases} 1 & \text{for } j = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.17}$$

where Y_r denotes the number of arrivals during the rth service time of length S_r.

Also using the backward Chapman Kolmogorov equation for $r = n-1, \quad 2$ and $j = 1, \quad n+1-r,$

$$\begin{aligned}
{}_0p_{j\,0}^{r-1\,n} &= Pr\{A^{r\,n-1}\ N_n = 0|N_{r-1} = j\} \\
&= \sum_{l=max(1\,j-1)}^{n-r} Pr\{N_r = l|N_{r-1} = j\}\ Pr\{A^{r+1\,n-1}\ N_n = 0|N_r = l\} \\
&= \sum_{l=max(1\,j-1)}^{n-r} Pr\{Y_r = l+1-j\}\ {}_0p_{l\,0}^{r\,n} \qquad (2\ 18)
\end{aligned}$$

and

$${}_0p_{j\,0}^{n-1\,n} = \delta_{1j}\ Pr\{Y_n = 0\} \qquad (2\ 19)$$

All other terms ${}_0p_{j\,1}^{r-1\,n-1}$ are 0 from (2 15)

The queue length estimate at time t_r $t_0 \le t \le t_n$ is then given by

$$N_r = \sum_{j=0}^{\infty} j p_{j|n}^{r} \qquad (2\ 20)$$

For an n customer busy period the computational complexity for an exact solution is $O(n^3)$ This can be reduced to $O(n^2 \log n)$ for an approximate solution based on a Gaussian approximation to a bound for N_r

Daley and Servi [DaS92] also give an $O(n)$ algorithm for an online estimate of the queue length at any departure instant inside a busy period If only $(t_0\ t_1 \quad t_r)$ are given and no information about events after time t_r is given the probability that the queue length at time t_r is j equals

$$\frac{{}_0p_{1\,j}^{0\,r}}{\sum_{l=1}^{\infty} {}_0p_{1\,l}^{0\,r}} \qquad (2\ 21)$$

which can be used to compute the online estimate of the queue length at time t_r

These results are essentially a rederivation of the results in [BeS92]

2 4 Feasible Arrival Vector Manjunath and Molle

Manjunath and Molle [MaM96] proposed passive estimation algorithms for queueing delays in LANs and polling systems A brief description of these results follow

Consider an interval $\mathcal{I} = (0, t_n]$ that includes n customer arrivals from a Poisson source, subject to the constraints that the mth arrival occurred no later than time t_m,

$m = 1, \ldots, n$. Let τ_m be the actual arrival time for the mth customer. Define $\tau_0 \triangleq 0$. For a single server queue $\mathcal{I}$ represents a busy period. The interval $\mathcal{I}$ and the constraints t_m are assumed to be known. Define the vectors $\underline{t} = [t_1 \; t_2 \; \ldots \; t_n]$ and $\underline{\tau} = [\tau_1 \; \tau_2 \; \ldots \; \tau_n]$.

Now if n arrivals from a Poisson source occurred within the interval $\mathcal{I}$, the joint density of their arrival times, $f'_n(\underline{\tau})$ is given by

$$f'_n(\underline{\tau}) = \frac{n!}{(t_n)^n} \qquad (2.22)$$

This allows the first arrival to occur within the interval $\mathcal{I}$ rather than defining its left hand boundary.

The arrivals are subject to the constraint that the mth arrival occurred before t_m. Thus, for any feasible arrival vector $\underline{\tau}$, the corresponding joint density $f_n(\underline{\tau})$ is obtained from $f'_n(\underline{\tau})$ by multiplying it by the normalization constant $1/p_n$, where p_n is the probability that a randomly chosen arrival vector would satisfy these constraints. p_n is given by

$$p_n = \int_{\tau_1=0}^{t_1} \int_{\tau_2=\tau_1}^{t_2} \ldots \int_{\tau \;=\tau_{n-1}}^{t_n} \frac{n!}{(t_n)^n} \, d\tau_n \ldots d\tau_2 \, d\tau_1 \qquad (2.23)$$

Therefore, the joint density of the arrival times given the constraint vector is given by

$$f_n(\underline{\tau}) = \begin{cases} \frac{f_n(\underline{\tau})}{p} & \text{for } i = 1, 2, \ldots \\ 0 & \text{for } i = 0 \end{cases} \qquad (2.24)$$

As long as $\underline{\tau}$ satisfies the given constraints, i.e. $\tau_{i-1} < \tau_i < t_i$ for all $i = 1, 2, \ldots, n$, otherwise $f_n(\underline{\tau}) = 0$.

Define $\phi_{p,q}(\tau_p)$ for $q \geq p$ and $p \geq 0$ as

$$\phi_{p,q}(\tau_p) \;=\; \begin{cases} 1 & \text{for } p = q \\ \int_{\tau_{p+1}=\tau_p}^{t_{p+1}} \int_{\tau_{p+2}=\tau_{p+1}}^{t_{p+2}} \ldots \int_{\tau_q=\tau_{q-1}}^{t_q} d\tau_q \ldots d\tau_{p+1} & \text{for } q > p \end{cases} \qquad (2.25)$$

It can be shown that $\phi_{p,q}$ is a polynomial of degree $q-p$ in τ_p and can be represented by

$$\phi_{p,q}(\tau_p) = \sum_{j=0}^{q-p} c_{p,q}(j) \; \tau_p^j \qquad (2.26)$$

where

$$\left. \begin{aligned} c_{p,q+1}(0) &= \textstyle\sum_{j=0}^{q-p} \; c_{p+1,q+1}(j) \; \frac{t_{p+1}^{j+1}}{j+1} \\ c_{p,q+1}(j) &= -\frac{c_{p+1,q+1}(j-1)}{j} \end{aligned} \right\} \qquad (2.27)$$

After further manipulations p_n and $f_n(\underline{\tau})$ reduce to

$$
\begin{aligned}
p_n &= \frac{n!}{(t_n)^n}\,\phi_{0\,n}(0) && (2\ 28)\\
f_n(\underline{\tau}) &= \frac{1}{\phi_{0\,n}(0)} && (2\ 29)
\end{aligned}
$$

The minimum mean square error estimate $E\tau_m$, of the arrival time of the mth cus tomer out of n in the interval $\mathcal{I}$ can be computed as given below

$$
\begin{aligned}
E\tau_m &= \int_{\tau_1=0}^{t_1} \quad \int_{\tau\ =\tau\ -1}^{t} \tau_m\, f_n(\underline{\tau})\, d\tau_n \quad d\tau_1\\
&= \frac{1}{\phi_{0\,n}(0)} \int_{\tau_1=0}^{t_1} \quad \int_{\ =\tau\ -1}^{t} \tau_m\, d\tau_n \quad d\tau_1
\end{aligned}
$$

This simplifies to

$$
E\tau_m = \frac{1}{c_{0\,n}(0)} \sum_{j=1}^{n-m+1} c_{m\,n}(j-1)\left[\sum_{i=1}^{m}(-1)^{i+1}\, t_{m+1-i}^{j+i}\, c_{0\,m-i}(0)\, \frac{j!}{(j+i)!}\right] \qquad (2\ 30)
$$

Now the waiting time of the mth customer in an n customer busy period is

$$
w_m = t_m - \tau_n \qquad (2\ 31)
$$

Taking expectations, we get

$$
Ew_m = t_m - E\tau_m \qquad (2\ 32)
$$

which may be computed using (2 30) Then the estimated queue length at the mth departure point is the difference between the estimated number of arrivals and the actual number of departures upto that point, namely m

2 5 Estimating Queueing Using Cumulative Departure Count Information

The collection of data pertaining to the service initiation and completion times for all the customers seived in a queueing system e g a node in a data network, amounts to a significant storage effoit on the pait of the monitoring agent The sheer volume of traf fic handled by the nodes limits the utility of queue inferencing algorithms that require extensive transactional data Hence algorithms requiring less frequently collected data assume importance for real networks Such data may, however, be less informative The cumulative number of departures in a given interval can be easily obtained from network management agents We intend to use existing queue inferencing algorithms to estimate the customer waiting times in a single server queueing system using the cumulative departure count information and the arrival rate information The development of this queue inferencing method for an M/M/1 queue is discussed below

The state of the server in a queue alternates between idle when the system is empty and busy , when a customer is receiving service at the server For a single server system the idle' period of the server corresponds to the idle period of the queueing system and the period during which the server is providing service to customers constitutes a busy period of the system The busy periods and the idle periods alternate in a queueing system An idle period and an adjacent busy period constitute a cycle of the system We divide the polling interval into cycles of steady state average length comprising of steady state average length idle and busy periods since these are the minimum mean square error estimates (*MMSE*) We distribute the total departures uniformly among these busy periods Thus we generate the kind of data as required by existing queue inferencing algorithms, and use these algorithms to estimate the waiting times of the customers We will use the algorithm of Manjunath and Molle [MaM96] for a single server FCFS queue in our scheme In the next section we will justify our choice of a uniform distribution for the total departures among the *MMSE* cycles

2 5 1 Distribution of Departures among the "Busy Periods"

We want a method to distribute the cumulative departures N, that occurred in a polling interval among the *MMSE* 'busy periods' Let the polling interval be divided into k cycles, and hence k busy periods Let us denote by n the number of customers allocated to the ith busy period Let us define $\underline{n} = [n_1, n_2, \quad n_k]$ The distribution of customers among the k busy periods is subject to the constraint that the sum of the customers allocated to all (i e k) busy periods must be equal to the total number of departures i e Φ N

Recognizing that but for the above constraint, the n_i s are iid random variables we have a situation similar to a product form queueing network Therefore

$$Pr\{\underline{n} \mid \sum_{i=1}^{k} n_i = N\} = \frac{f(n_1)f(n_2) \quad f(n_k)}{Pr\{\sum_{i=1}^{k} n_i = N\}} \tag{2 33}$$

where $f(n_i)$ is the probability mass function for the number of customers served in the ith busy period From [KLEIN](p 218) $f(n_i)$ is given by

$$f(n_i) = \frac{1}{n_i} \begin{pmatrix} 2n_i - 2 \\ n_i - 1 \end{pmatrix} \rho^{n_i - 1} (1+\rho)^{1-2n_i} \tag{2 34}$$

and

$$\rho = \frac{\lambda}{\mu} \tag{2 35}$$

Since n_i s are iid random variables

$$\begin{aligned} Pr\{\sum_{i=1}^{k} n_i = N\} &= \underbrace{f(n) \otimes f(n) \otimes \quad \otimes f(n)}_{k \text{ times}} \mid_{n=N} \\ &= f^{(k)}(n) \mid_{n=N} \end{aligned} \tag{2 36}$$

Therefore from (2 33) and (2 36) we have

$$Pr\{\underline{n} \mid \sum_{i=1}^{k} n_i = N\} = \frac{f(n_1)f(n_2) \quad f(n_k)}{f^{(k)}(n) \mid_{n=N}} \tag{2 37}$$

Thus we see that the probability mass function for $\underline{n}$ is independent of any permutation of a given instance of $\underline{n}$ The symmetry of the joint probability mass function points to the uniform distribution of customers to all the busy periods as being the mean case So a uniform distribution i e $n_i = (N/k)$ for $i = 1\ 2$ will minimize the variance

2 5 2 Numerical Results

In the proposed scheme we divide the polling interval in MMSE cycles consisting of MMSE idle periods and adjacent MMSE busy periods We then distribute the total departures in that interval uniformly among the busy periods to generate the data required by the existing queue inferencing techniques We estimate the waiting times using the queue inferencing algorithm proposed by Manjunath and Molle [MaM96]

We carried out simulation of an M/M/1 FCFS queue for the purpose of testing the proposed scheme Tests were run using polling intervals of different sizes (100 200 500 1000

μ^{-1} units) for average arrival rate (λ) ranging from 0.1 to 0.9 (normalized w.r.t. the average service rate μ). Time is measured in units of average service time μ^{-1}. The estimated waiting times are compared with the actual values and the bias in estimation is computed. It is observed that the scheme is biased towards giving a lower estimate. The results are presented in tables 2.1 - 2.4. It may be observed from these tables and fig. 2.1 that the proposed scheme is biased towards lower estimates which worsen with increasing λ. It may also be noted that the performance is unaffected by the choice of polling interval.

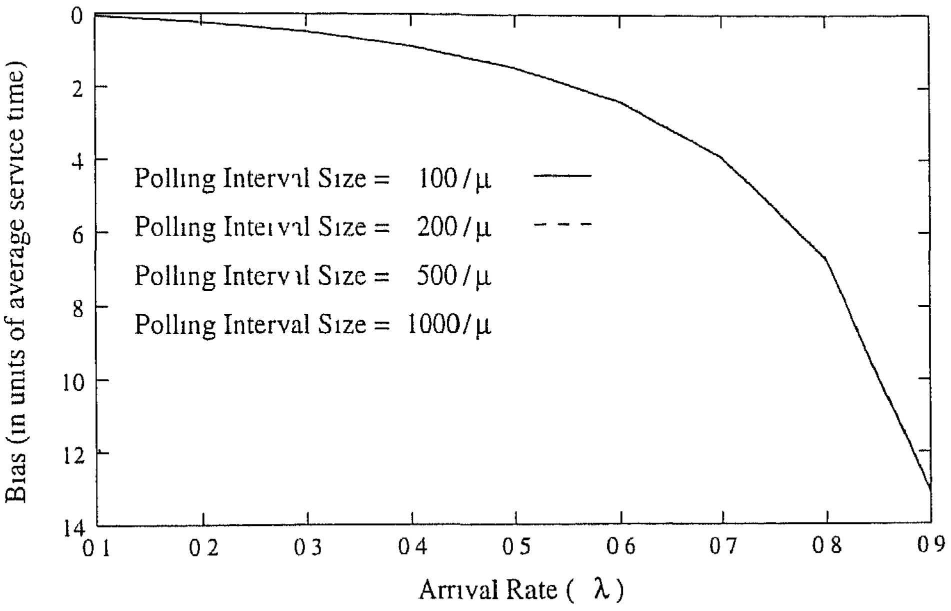

Figure 2.1 Variation of bias with λ

2.5.3 "Offset" at the Polling Edge

Our scheme assumes that given a proper choice of polling interval size T, we can have exactly $k = T/\bar{c}$ busy periods inside the polling interval. This implicitly assumes that the leading edge of the polling interval falls in an idle period. Let the distance of the leading polling interval edge from the beginning of the subsequent busy period be called as the "offset". If the leading polling interval edge falls in a busy period, the "offset" is the difference between the beginning of the current busy period and the polling edge and is taken to be negative in this case (fig. 2.2). Let us denote the "offset" by β. Then

λ	Mean Estimated Waiting Time	Bias
0 100000	0 049971	0 061862
0 200000	0 097838	0 226910
0 300000	0 170402	0 449355
0 400000	0 289109	0 799211
0 500000	0 469522	1 390885
0 600000	0 743035	2 517193
0 700000	1 146156	3 990822
0 800000	1 736006	7 511800
0 900000	2 857449	11 292483

Table 2 1 Performance of scheme for polling interval size = 100 μ^{-1}

λ	Mean Estimated Waiting Time	Bias
0 100000	0 042382	0 082124
0 200000	0 091530	0 239443
0 300000	0 166029	0 461633
0 400000	0 286228	0 805425
0 500000	0 468496	1 394747
0 600000	0 742648	2 516473
0 700000	1 146860	3 988634
0 800000	1 738224	7 497245
0 900000	2 860715	11 278063

Table 2 2 Performance of scheme for polling interval size = 200 μ^{-1}

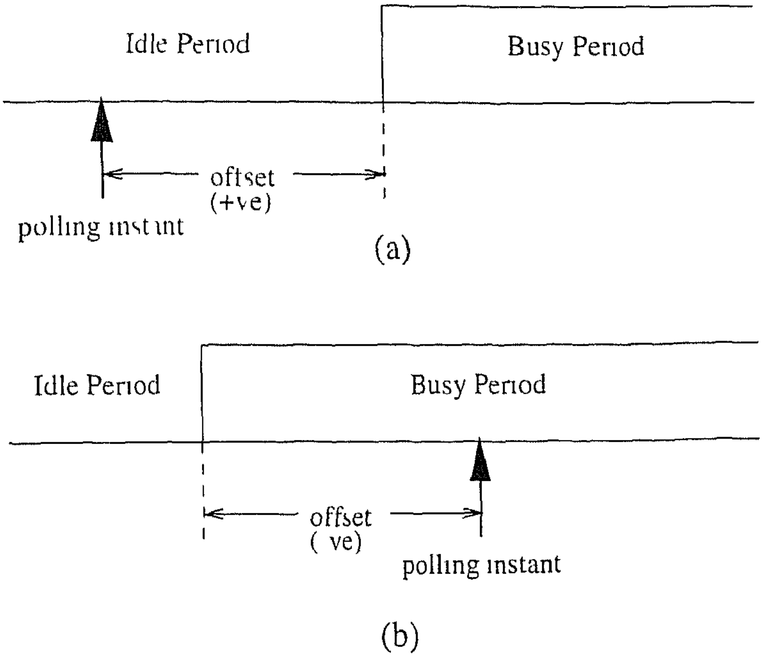

Figure 2.2 Offset at the polling edge

we have

$$\beta = Pr\{idle\}\frac{1}{\lambda} - Pr\{busy\}\frac{\overline{Y^2}}{2(\overline{Y})} \tag{2.38}$$

where λ = average arrival rate to the system

$\overline{Y}$ = mean busy period length

$\overline{Y^2}$ = second central moment of the busy period length

For an M/M/1 for $\mu = 1$ the values of $\overline{Y}$ and $\overline{Y^2}$ are given by the following expressions

$$\overline{Y} = \frac{1}{(1-\lambda)} \tag{2.39}$$

$$\overline{Y^2} = \frac{2}{(1-\lambda)^3} \tag{2.40}$$

Also we have

$$Pr\{idle\} = 1-\lambda \tag{2.41}$$

$$Pr\{busy\} = \lambda \tag{2.42}$$

Hence we have,

$$\begin{aligned} \beta &= \frac{(1-\lambda)}{\lambda} - \frac{\lambda}{(1-\lambda)^2} \\ &= \frac{(1-\lambda)^3-\lambda^2}{\lambda(1-\lambda)^2} \end{aligned} \tag{2.43}$$

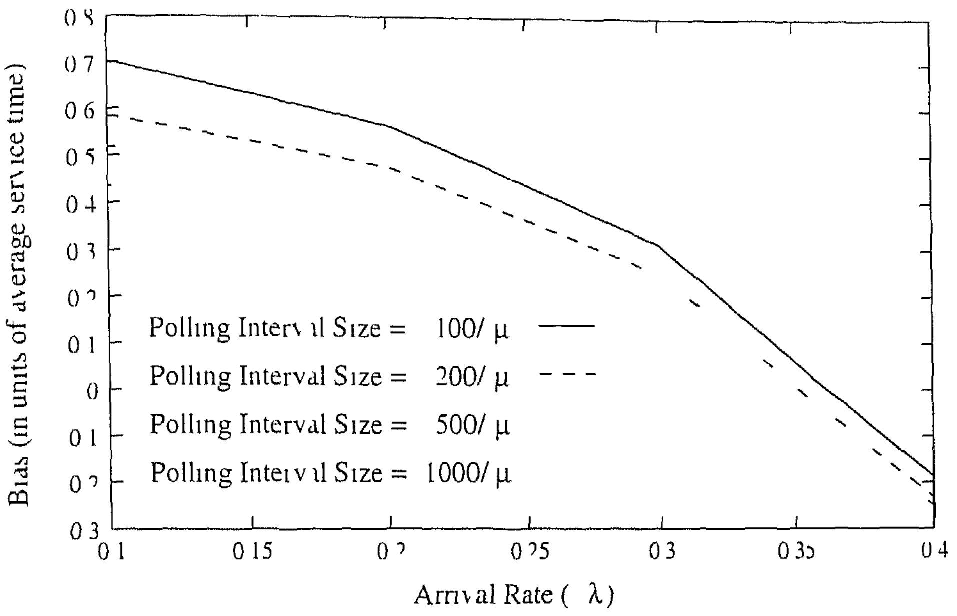

Figure 2.3 Bias Performance after offset correction

Thus we note that the mean offset of the polling edge is non zero. This is contrary to our assumption that the offset is zero for all values of λ. We tried to account for this problem in our scheme but were unable to find a suitable way of introducing the necessary correction in case of negative offsets. fig 2.3 shows the bias in the results for the cases where the offset is non negative. The polling intervals chosen are $100\mu^{-1}$, $200\mu^{-1}$, $500\mu^{-1}$ and $1000\mu^{-1}$. We can note the improvement in bias performance with this scheme as compared to the earlier method. But this method cannot be used with higher values of λ because the associated negative offset cannot be corrected.

2.6 Discussion

In this chapter, we discussed the early work on queue inferencing techniques. We formulated a queue inferencing scheme to obtain waiting times requiring cumulative departure counts. This scheme uses the information of the departure counts to generate the kind of information required by existing queue inferencing techniques. Existing queue inferencing algorithms are then applied on this generated data to obtain waiting time estimates. We checked the scheme by the way of simulation and found out that the scheme is biased towards giving smaller estimates than the real values. We presented a scheme to account for the offset encountered by the polling edge. We observed that this scheme has a better error performance but we were unable to find a way of applying this offset correction for the cases where the offset is negative

Chapter 3

Residual Queue Length Estimation

3 1 Introduction

The information about the number of departures occurring in a specified time interval is easy to obtain for a queueing system This information coupled with the knowledge of arrival rate may be used to extract queue length estimates for the system at the end of the interval specified We will call the queue length at the end of such an interval as the residual' queue length In this chapter, we present a method to estimate the residual queue length conditioned on the number of departures in a specified time interval and the arrival rate to the queueing system

In section 3 2 we discuss the approach to obtain the residual queue lengths for an M/M/1 queue In section 3 3, we derive the joint distribution for the departures and the queue length In section 3 4, we derive an formula for estimating residual queue lengths Our algorithm assumes the knowledge of arrival rate In section 3 5, we discuss the error involved in the measurement of arrival rates Finally we present some numerical results in section 3 6

3 2 Residual Queue Length Estimation Theory

Consider an M/M/1 queue Let the system be in the idle state at some time t_0 Let us denote the residual queue length at time $t_0 + t$ by Q_{t_0+t} Let us denote the departures in the interval by D_{t_0+t} Let $D_{t_0+t} = d$ Without loss of generality, we can define $t_0 = 0$

Then we have $Q_t = Q_{t_0+t}$ and $D_t = D_{t_0+t} = d$ By the law of total probability we have

$$Pr\{Q_t = i | D_t = d\} = \frac{Pr\{Q_t = i \; D_t = d\}}{Pr\{D_t = d\}} \tag{3 1}$$

Here $Pr\{Q_t = i | D_t = d\}$ is the probability of the residual queue length being i conditioned on the number of departures being d in interval $(0 \; t]$ $Pr\{Q_t = i \; D_t = d\}$ is the joint probability for the event $Q_t = i$ and $D_t = d$ $Pr\{D_t = d\}$ is the probability of d departures occurring in the time interval $(0 \; t]$ Then the estimate for the residual queue length is given by

$$\begin{aligned} E\{Q_t | D_t = d\} &= \sum_{i=0}^{\infty} i Pr\{Q_t = i | D_t = d\} \\ &= \sum_{i=0}^{\infty} i \frac{Pr\{Q_t = i \; D_t = d\}}{Pr\{D_t = d\}} \end{aligned} \tag{3 2}$$

The probability of d departures occurring in a time interval of size t for an M/M/1 queue with arrival rate λ is given by [COHEN](pp 199 200) for $d = 0 \; 1 \quad t \geq 0$ as

$$Pr\{D_t = d\} = \frac{(\lambda t)^d e^{-\lambda t}}{d!} \tag{3 3}$$

The joint probability for D_t and Q_t is obtained in the next section

3 3 Joint Density Function for Departures and Residual Queue Length

Consider an M/M/1 queue where the arrival rate normalized to the service rate is λ D_t is the number of departures during the interval $(0 \; t]$ Let us define $D_0 = 0$ Let X_t be the system state at time t Let us define $X_0 = 0$ i e the system is considered empty at the beginning of time interval $(0 \; t]$ Let us define

$$H_{ij}(t) = Pr\{X_t = i \; D_t = j\} \tag{3 4}$$

Let us define for $|z_1| \leq 1$, $|z_2| \leq 1$, **Re** $s > 0$

$$h(s, z_1, z_2) \triangleq \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} {z_1}^i {z_2}^j \int_0^{\infty} e^{-st} H_{ij}(t) dt \tag{3 5}$$

$$h_{ij}(s) \triangleq \int_0^{\infty} e^{-st} H_{ij}(t) dt \tag{3 6}$$

The expression for $h(s\ z_1\ z_2)$ is [COHEN](pp 197 198)

$$h(s\ z_1\ z_2) = \frac{(z_2 - z_1)\sum_{j=0}^{\infty} h_{0j}(s)\, z_2^j - z_1}{\lambda z_1^2 - (1+\lambda+s)z_1 + z_2} \tag{3 7}$$

The denominator of (3 7) may be treated as a quadratic in z_1 Let us denote the zeros of the denominator by $x_1(z_2)$ and $x_2(z_2)$ so that

$$x_1(z_2) \triangleq \frac{(1+\lambda+s) + \sqrt{(1+\lambda+s)^2 - 4\lambda z_2}}{2\lambda} \tag{3 8}$$

$$x_2(z_2) \triangleq \frac{(1+\lambda+s) - \sqrt{(1+\lambda+s)^2 - 4\lambda z_2}}{2\lambda} \tag{3 9}$$

It can be proved by Rouche s theorem that for **Re** $s > 0$ $|z_2| \leq 1$ [COHEN](p 198)

$$|x_1(z_2)| > 1,\ |x_2(z_2)| < 1$$

Since $h(s\ z_1\ z_2)$ should be an analytic function of z_1 inside the unit circle for fixed s z_2 with **Re** $s > 0$ $|z_2| \leq 1$ it follows that $x_2(z_2)$ should also be a zero of the numerator of (3 7) Hence for $|z_2| \leq 1$ **Re** $s > 0$

$$\sum_{j=0}^{\infty} h_{0j}(s) z_2^j = \frac{x_2(z_2)}{z_2 - x_2(z_2)} \tag{3 10}$$

$$h(s\ z_1\ z_2) = \frac{(z_2 - z_1)x_2(z_2) - (z_2 - x_2(z_2))z_1}{(z_2 - x_2(z_2))\left\{\lambda z_1^2 - (1+\lambda+s)z_1 + z_2\right\}} \tag{3 11}$$

We note that

$$\lambda z_1^2 - (1+\lambda+s)z_1 + z_2 = \lambda(z_1 - x_1(z_2))(z_1 - x_2(z_2)) \tag{3 12}$$

Hence in (3 11) we get

$$\begin{aligned} h(s\ z_1, z_2) &= \frac{z_2(x_2(z_2) - (z_1))}{(z_2 - x_2(z_2))\lambda(z_1 - x_1(z_2))(z_1 - x_2(z_2))} \\ &= \frac{-z_2}{(z_2 - x_2(z_2))(z_1 - x_1(z_2))\lambda} \\ &= \left\{1 - \frac{x_2(z_2)}{z_2}\right\}^{-1} \frac{1}{\lambda x_1(z_2)} \left\{1 - \frac{z_1}{x_1(z_2)}\right\}^{-1} \end{aligned} \tag{3 13}$$

We choose s such that $|x_2(z_2)| < |z_2|$ We also know that $|z_1| \leq 1 \leq |x_1(z_2)|$ Expanding the right hand side of (3 13), we get

$$h(s, z_1\ z_2) = \sum_{l=0}^{\infty}\left\{\frac{x_2(z_2)}{z_2}\right\}^l \frac{1}{\lambda x_1(z_2)} \sum_{m=0}^{\infty}\left\{\frac{z_1}{x_1(z_2)}\right\}^m \tag{3 14}$$

From (3 8) and (3 9) we have

$$x_1(z_2)x_2(z_2) = \frac{z_2}{\lambda} \qquad (3\ 15)$$

From (3 14) and (3 15) we get

$$h(s, z_1, z_2) = \sum_{l=0}^{\infty}\left\{\frac{x_2(z_2)}{z_2}\right\}^{l}\frac{x_2(z_2)}{z_2}\sum_{n=0}^{\infty}\left\{\frac{\lambda z_1 x_1(z_2)}{}\right\} \qquad (3\ 16)$$

$$= \sum_{m=0}^{\infty}\lambda^m z_1{}^m\sum_{l=0}^{\infty}\left\{\frac{x_2(z_2)}{z_2}\right\}^{l+m+1} \qquad (3\ 17)$$

Now $x_2{}^n(z_2)$ is given by the following relation [COHEN](p 198) for $n = 1\ 2$

$$x_2{}^n(z_2) = \sum_{r=0}^{\infty} z_2{}^{n+r}\int_0^{\infty} e^{-(1+\lambda+s)t}\frac{n\lambda^r}{r!(n+r)!}t^{2r+n-1}dt \qquad (3\ 18)$$

Hence we have

$$h(s, z_1, z_2) = \sum_{m=0}^{\infty}\lambda^m z_1{}^m\sum_{l=0}^{\infty}\sum_{r=0}^{\infty} z_2{}^r\int_0^{\infty} e^{-(1+\lambda+s)t}\frac{(l+m+1)\lambda^r t^{2r+l+m}}{r!(l+m+r+1)!}dt \qquad (3\ 19)$$

This may be inverted with respect to s, z_1 and z_2 (ref Appendix A) We then have $H_{ij}(t)$ given by

$$H_{ij}(t) = \frac{e^{-\lambda t}\lambda^{i+j}t^{j-1}}{j!}\left[(t-j)\left\{1-e^{-t}\sum_{m=0}^{i+j}\frac{t^m}{m!}\right\}+te^{-t}\frac{t^{i+j}}{(i+j)!}\right] \qquad (3\ 20)$$

Recognizing that H_{ij} is the joint probability density for departures and the residual queue length we get

$$Pr\left\{Q_t = i\ D_t = j\right\} = H_{ij}(t) \qquad (3\ 21)$$

From (3 20) and (3 21) we have

$$Pr\left\{Q_t = i\ D_t = j\right\} = \frac{e^{-\lambda t}\lambda^{i+j}t^{j-1}}{j!}\left[(t-j)\left\{1-e^{-t}\sum_{m=0}^{i+j}\frac{t^n}{m!}\right\}+te^{-t}\frac{t^{i+j}}{(i+j)!}\right] \qquad (3\ 22)$$

3 4 The Estimate of the Residual Queue Length

With the help of the joint density function for departures and residual queue length for an M/M/1 queue (3 22) and (3 3) we have

$$\begin{aligned} Pr\left\{Q_t = i|D_t = d\right\} &= \frac{d!}{e^{-\lambda t}(\lambda t)^d}\frac{e^{-\lambda t}\lambda^{i+d}t^{d-1}}{d!}\left[(t-d)\left\{1-e^{-t}\sum_{n=0}^{i+d}\frac{t^m}{m!}\right\}\right. \\ &\qquad \left. +te^{-t}\frac{t^{i+d}}{(i+d)!}\right] \\ &= \frac{\lambda^i}{t}\left[(t-d)\left\{1-e-t\sum_{m=0}^{i+d}\frac{t_m}{m!}\right\}+e^{-t}\frac{t^{i+d+1}}{(i+d)!}\right] \qquad (3\ 23) \end{aligned}$$

Then the estimate for residual queue length at the end of an observation interval of size t is given by the following relation

$$
\begin{aligned}
E\{Q_t|D_t = d\} &= \sum_{i=0}^{\infty} i P\{Q_t = i|D_t = d\} \\
&= \sum_{i=0}^{\infty} i \frac{\lambda^i}{t}\left[(t-d)\left\{1 - e-t \sum_{n=0}^{i+d} \frac{t^m}{m!}\right\} + e^{-t}\frac{t^{i+d+1}}{(i+d)!}\right] \\
&= \sum_{i=0}^{\infty} \frac{(t-d)}{t} i\lambda^i - \sum_{i=0}^{\infty} \frac{(t-d)e^{-t}}{t} i\lambda^i \sum_{m=0}^{i+d} \frac{t^m}{m!} \\
&\quad + \sum_{i=0}^{\infty} e^{-t} i\lambda^i \frac{t^{i+d+1}}{(i+d)!}
\end{aligned}
\qquad (3.24)
$$

The above equation can be solved (ref. Appendix A) to obtain an expression for $E\{Q_t|D_t = d\}$ is given below

$$
\begin{aligned}
E\{Q_t|D_t = d\} &= \frac{(t-d)\lambda}{(1-\lambda)^2 t} - \frac{(t-d)e^{-(1-\lambda)t}}{(1-\lambda)\lambda^{d-1}} - \frac{(t-d)e^{-(1-\lambda)t}}{(1-\lambda)^2\lambda^d t} + \frac{te^{-(1-\lambda)t}}{\lambda^{d-1}} - \frac{de^{-(1-\lambda)t}}{\lambda^d} \\
&\quad + \frac{(t-d)(d+1)}{(1-\lambda)\lambda^d t} e^{-(1-\lambda)t} + \frac{(t-d)e^{-t}}{(1-\lambda)\lambda^{d-1}}\phi(\lambda t, d-2) - \frac{te^{-t}}{\lambda^{d-1}}\phi(\lambda t, d-2) \\
&\quad + \frac{(t-d)e^{-t}}{(1-\lambda)^2\lambda^d t}\phi(\lambda t, d-1) - \frac{(t-d)e^{-t}(d+1)}{(1-\lambda)\lambda^d t}\phi(\lambda t, d-1) \\
&\quad + \frac{de^{-t}}{\lambda^d}\phi(\lambda t, d-1) + \frac{(t-d)e^{-t}(d+1)}{(1-\lambda)t}\phi(t, d-1) \\
&\quad - \frac{(t-d)de^{-t}}{(1-\lambda)t}\phi(t, d-1) - \frac{(t-d)e^{-t}}{(1-\lambda)^2 t}\phi(t, d-1)
\end{aligned}
\qquad (3.25)
$$

where $\phi(a, n)$ is defined as

$$
\phi(a, n) \triangleq \sum_{r=0}^{n} \frac{a^r}{r!} \qquad (3.26)
$$

It may be noted that the given expression requires $O(d)$ calculations

3.5 Traffic Arrival Rate Estimation

The algorithm derived in (3.25) assumes the knowledge of normalized traffic arrival rate. An obvious way of obtaining this information is to measure the utilization of the server. This is easily obtained by observing the fraction of the time in some interval $(0, t)$ that the server is busy. The length for which this observation is to be carried out, i.e. the value of t depends on the accuracy with which we want to obtain λ. In [Kuma92] Kumar shows that the error in the estimate of $(1-\lambda)$ obtained from this observation is bounded

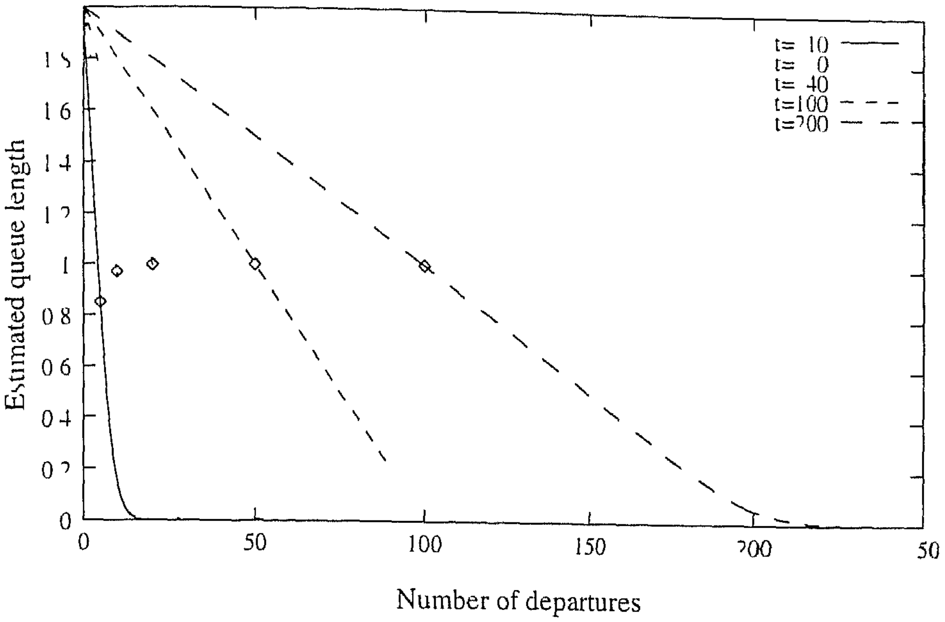

Figure 3.1 Estimated Queue Length for $\lambda = 0.5$ Points marked for $D = \lambda t$

by

$$\frac{1}{t}\left|\frac{1}{1-\lambda_1}-\frac{1}{1-\lambda_2}\right| \tag{3.27}$$

where λ_1 is the arrival rate before the observation began and λ_2 is the actual arrival rate in the current window

3.6 Numerical Results

Equation (3.25) was used to estimate the residual queue length $Q_{t|D_t}$ for various values of the arrival rate (λ), and for various values of time interval (t) $Q_{t|D_t}$ as a function of d for various values of t is shown in figures 3.1 3.4 It may be observed that as the value of t increases the estimated queue length for $d = \lambda t$ approaches the steady state value of $\frac{\lambda}{(1-\lambda)}$, with the arrival rate normalized to the service rate It may also be seen that $Q_{t|D=\lambda t}$ approaches the steady state average value faster for smaller λ This may be observed from table 3.1 also where we have tabulated the values of $Q_{t|D=\lambda t}$ for $d = \lambda t$ for different values of λ and t

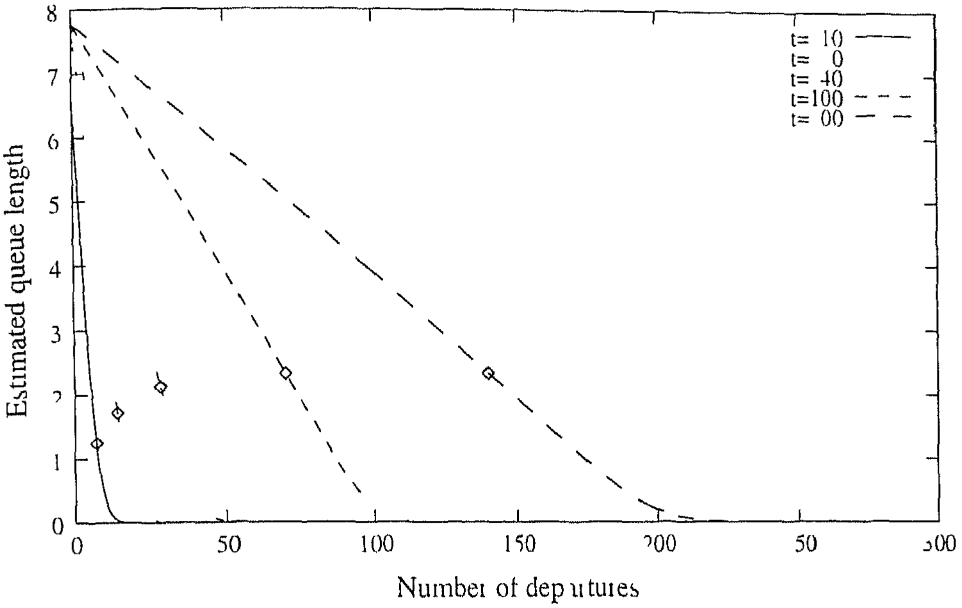

Figure 3 2 Estimated Queue Length for $\lambda = 0\ 7$ Points marked for $D = \lambda t$

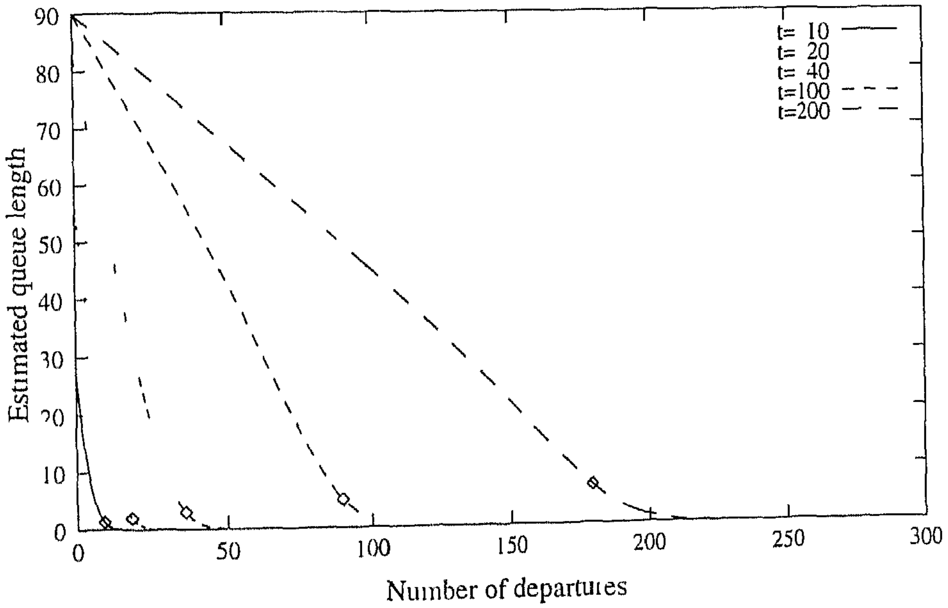

Figure 3 3 Estimated Queue Length for $\lambda = 0\ 9$ Points marked for $D = \lambda t$

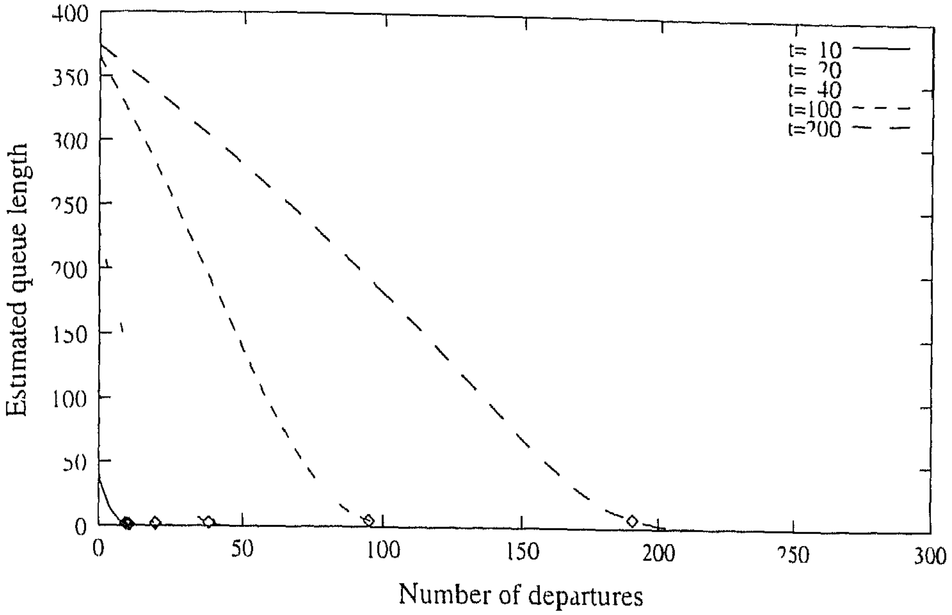

Figure 3 1 Estimated Queue Length for $\lambda = 0\ 95$ Points marked for $D = \lambda t$

Time	Estimated Queue Length		
(in μ^{-1} units)	$\lambda = 0\ 5$	$\lambda = 0\ 7$	$\lambda = 0\ 9$
10	0 850110	1 253427	1 352603
20	0 969804	1 722327	2 014822
40	0 998679	2 124539	2 997568
100	1 000000	2 323942	4 925797
200	1 000000	2 333275	6 724913
1000	1 000000	2 333333	8 999621
$\frac{\lambda}{1-\lambda}$	1 000000	2 333333	9 000000

Table 3 1 The variation of $Q_{t\mid D=\lambda t}$ with t

3.7 Discussion

In this chapter we presented a new approach to obtain residual queue length estimates using the cumulative departure count information. We derived an expression for the joint probability distribution of the residual queue length and the cumulative departure count and derived our estimation formula using the law of total probability. The complexity of our estimation algorithm is $O(d)$ where d is the cumulative departure count. We discussed a technique for estimation of arrival rate and the error involved therein. Finally we presented some numerical results for our algorithm.

Chapter 4

Conclusions and Future Work

4 1 Conclusions

In this thesis we presented queue inferencing methods which use the easily obtainable cumulative departure information for estimating queueing parameters

First we proposed a scheme to estimate customer waiting times in an M/M/1 queue The proposed scheme divides a polling interval in cycles of steady state average size and uniformly divides the total departures in the polling interval among the cycles We discovered that the scheme is biased towards making too small estimates which worsened with the increase in the customer arrival rate We proposed a correction to the above scheme using the observation that the leading edge of the polling interval does not always fall in an idle period We discovered that the bias performance improves after this correction but we could not formulate a way of applying this correction for higher arrival rate queues

Next we developed an $O(d)$ algorithm for estimating the residual queue lengths using the cumulative departure count d We derived the joint probability distribution of the residual queue length and the cumulative departure count for a given time interval and used the fact that the departure process for an M/M/1 queue is also Poisson We derived our result using the law of total probability Our algorithm assumes the knowledge of customer arrival rate We discussed the technique for the measurement of server utilization factor, and the approximation involved We also presented some numerical results

In the next section, we will discuss the possible use of PING for queue inferencing which is a new direction for future work.

4.2 Queue Inferencing using PING

PING, or Packet InterNet Groper, is a facility supported by TCP/IP networks to check whether a remote system is alive or not. PING sends constant (adjustable) size packets from the querying node to the queried node. The queried node then echoes the packets back to the querying node. PING packets carry the the timestamp of the transmission time. This is used to calculate the round trip time. This round-trip time is due to queueing delays at both the nodes and the delay experienced at the intermediate nodes. Our interest would be to find out how this roundtrip time information provided by the PING packet can be used to estimate the delays experienced by regular network traffic.

The system described above can be modelled as queueing system with two classes of customers. PING packets are a class of customers originating from a regular, possibly finite source. The network traffic will be the second class of customers with Poisson arrivals from an infinite source. It is obvious that the expected delays of the two classes of customers are different. This means that the information obtained from the RTT of the PING packet does not directly indicate the delays of the other traffic. We will need to use the RTT to estimate the delays experienced by the regular traffic by constructing the queueing model and obtaining the relation between the PING RTT and the queueing delays of the network traffic.

A queue inferencing scheme that uses the RTT information from PING can be modelled as a single server queue with two classes of customers: one from an infinite Poisson source, and other from a finite source which is fed back into the queue after a delay at the source. The steady state solution for such system is presented in [Box85, DoW87, BoC91]. These results may be used to develop this idea further.

Appendix A

A 1 Derivation of the Joint Density for Queue Length and the Number of Departures for an M/M/1 Queue

For an M/M/1 queue the value of $h(s, z_1, z_2)$ is given by [COHEN](pp 197 198)

$$h(s, z_1, z_2) = \frac{(z_2 - z_1)x_2(z_2) - (z_2 - x_2(z_2))z_1}{(z_2 - x_2(z_2))\{\lambda z_1^2 - (1+\lambda+s)z_1 + z_2\}} \qquad |z_1|, |z_2| \leq 1 \tag{A 1}$$

The roots of the quadratic term in the denominator of (A 1) are given by

$$x_1(z_2) = \frac{(1+\lambda+s)+\sqrt{(1+\lambda+s)^2-4\lambda z_2}}{2\lambda} \qquad |x_1(z_2)| \geq 1 \tag{A 2}$$

$$x_2(z_2) = \frac{(1+\lambda+s)-\sqrt{(1+\lambda+s)^2-4\lambda z_2}}{2\lambda} \qquad |x_2(z_2)| \leq 1 \tag{A 3}$$

Therefore by choosing s such that $|x_2(z_2)| < |z_2|$ we have

$$\begin{aligned} h(s, z_1, z_2) &= \frac{z_2(x_2(z_2) - z_1)}{(z_2 - x_2(z_2))\lambda(z_1 - x_1(z_2))(z_1 - x_2(z_2))} \\ &= \frac{z_2}{(z_2 - x_2(z_2))\lambda(x_1(z_2) - z_1)} \\ &= \sum_{m=0}^{\infty} \left\{\frac{x_2(z_2)}{z_2}\right\}^n \frac{1}{\lambda x_1(z_2)} \sum_{l=0}^{\infty} \left\{\frac{z_1}{x_1(z_2)}\right\}^l \end{aligned} \tag{A 4}$$

We observe that $x_1(z_2)x_2(z_2) = z_2/\lambda$ Therefore

$$\begin{aligned} h(s, z_1, z_2) &= \sum_{m=0}^{\infty} \left\{\frac{x_2(z_2)}{z_2}\right\}^m \frac{x_2(z_2)}{z_2} \sum_{l=0}^{\infty} \left\{\frac{z_1 x_2(z_2)\lambda}{z_2}\right\}^l \\ &= \sum_{l=0}^{\infty} z_1^l \lambda^l \sum_{m=0}^{\infty} \left\{\frac{x_2(z_2)}{z_2}\right\}^{l+m+1} \end{aligned} \tag{A 5}$$

For $r = 1, 2, \ldots$, [COHEN](p 198)

$$x_2^r(z_2) = \sum_{n=0}^{\infty} z_2^{r+n} \int_0^{\infty} e^{-(1+\lambda+s)t} \frac{r\lambda^n}{n!(r+n)!} t^{2n+r-1} dt \tag{A 6}$$

Hence we have

$$h(s, z_1, z_2) = \sum_{l=0}^{\infty} z_1^l \lambda^l \sum_{n=0}^{\infty} \sum_{n=0}^{\infty} z_2^n \int_0^{\infty} e^{-(1+\lambda+s)t} \frac{(l+m+1)\lambda^n}{n!(l+m+n+1)!} t^{2n+l+m} dt \quad (A.7)$$

Taking the inverse Laplace transform in (A.7) we get

$$h(t, z_1, z_2) = \sum_{l=0}^{\infty} z_1^l \lambda^l \sum_{n=0}^{\infty} \sum_{n=0}^{\infty} \frac{\lambda^n z_2^n (l+m+1)}{n!(l+m+n+1)!} e^{-(1+\lambda)t} t^{2n+l+m} \quad (A.8)$$

Taking the inverse Z transforms in (A.8) w.r.t. z_1 and z_2 we obtain

$$H_{i,d}(t) = e^{-(1+\lambda)t} \lambda^i \sum_{m=0}^{\infty} \frac{\lambda^d (m+i+1)}{d!(m+i+d+1)!} t^{2d+i+m} \quad (A.9)$$

Upon further simplification, (A.9) gives

$$H_{id}(t) = \frac{e^{-\lambda t} \lambda^{i+d} t^{d-1}}{d!} \left[(t-d) \left\{ 1 - e^{-t} \sum_{n=0}^{i+d} \frac{t^m}{m!} \right\} + te^{-t} \frac{t^{i+d}}{(i+d)!} \right] \quad (A.10)$$

A.2 Derivation of the Estimate of the Residual Queue Length for an M/M/1 Queue

The queue length at time t, conditioned on there being d departures in interval $[0, t)$ is given by

$$\begin{aligned} E[Q_t | D_t = d] &= \sum_{i=0}^{\infty} \frac{i\lambda^i}{t} \left[(t-d) \left\{ 1 - e^{-t} \sum_{m=0}^{i+d} \frac{t^m}{m!} \right\} + e^{-t} \frac{t^{i+d+1}}{(i+d)!} \right] \\ &= \sum_{i=0}^{\infty} \frac{i\lambda^i (t-d)}{t} - \sum_{i=0}^{\infty} \frac{i\lambda^i (t-d)}{t} e^{-t} \sum_{m=0}^{i+d} \frac{t^m}{m!} + \sum_{i=0}^{\infty} \frac{i\lambda^i}{t} e^{-t} \frac{t^{i+d+1}}{(i+d)!} \\ &= T_1 - T_2 + T_3 \qquad (A.11) \end{aligned}$$

Let us define $\phi(a, n)$ by

$$\phi(a, n) \triangleq \sum_{i=0}^{n} \frac{a^i}{i!} \quad (A.12)$$

We will now consider T_1, T_2 and T_3 separately.

$$\begin{aligned} T_1 &= \sum_{i=0}^{\infty} \frac{i\lambda^i (t-d)}{t} \\ &= \frac{(t-d)\lambda}{t} \sum_{i=0}^{\infty} i\lambda^{i-1} \end{aligned}$$

$$
\begin{aligned}
&= \frac{(t-d)\lambda}{t}\frac{1}{(1-\lambda)^2} \\
&= \frac{(t-d)\lambda}{(1-\lambda)^2 t} && \text{(A 13)}
\end{aligned}
$$

$$
\begin{aligned}
T_2 &= \sum_{=0}^{\infty} \frac{i\lambda^i (t-d)}{t} e^{-t} \sum_{m=0}^{i+d} \frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}}{t} \sum_{i=0}^{\infty} i\lambda \sum_{m=0}^{i+d} \frac{t^m}{m!} \\
&= \frac{(t-d)e-t}{t} \left[\sum_{=0}^{\infty} (i+d+1) \frac{\lambda^{i+d}}{\lambda^d} \sum_{m=0}^{i+d} \frac{t^m}{m!} - \sum_{i=0}^{\infty} (d+1) \frac{\lambda^{i+d}}{\lambda^d} \sum_{m=0}^{i+d} \frac{t^m}{m!} \right] \\
&= \frac{(t-d)e-t}{\lambda^d t} \sum_{i=0}^{\infty} (i+d+1)\lambda^{i+d} \sum_{m=0}^{+d} \frac{t^m}{m!} \\
&\quad - \frac{(t-d)e^{-t}(d+1)}{\lambda^d t} \sum_{=0}^{\infty} \lambda^{i+d} \sum_{m=0}^{i+d} \frac{t^m}{m!} \\
&= T_{2A} - T_{2B} && \text{(A 14)}
\end{aligned}
$$

We have

$$
\begin{aligned}
T_{2A} &= \frac{(t-d)e^{-t}}{\lambda^d t} \sum_{i=0}^{\infty} (i+d+1)\lambda^{i+d} \sum_{m=0}^{i+d} \frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}}{\lambda^d t} \sum_{k=d}^{\infty} (k+1)\lambda^k \sum_{m=0}^{k} \frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}}{\lambda^d t} \left[\sum_{k=0}^{\infty} (k+1)\lambda^k \sum_{m=0}^{k} \frac{t^m}{m!} - \sum_{k=0}^{d-1} (k+1)\lambda^k \sum_{m=0}^{t} \frac{t^m}{m!} \right] \\
&= T_{2A(a)} - T_{2A(b)} && \text{(A 15)}
\end{aligned}
$$

The term $T_{2A(a)}$ is given by

$$
\begin{aligned}
T_{2A(a)} &= \frac{(t-d)e^{-t}}{\lambda^d t} \sum_{k=0}^{\infty} (k+1)\lambda^k \sum_{m=0}^{k} \frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}}{\lambda^d t} \sum_{m=0}^{\infty} \sum_{k=m}^{\infty} (k+1)\lambda^k \frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}}{\lambda^d t} \sum_{m=0}^{\infty} \frac{[m(1-\lambda)\lambda^m + \lambda^m]}{(1-\lambda)^2} \frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}}{\lambda^d t} \sum_{m=0}^{\infty} \left[\frac{m\lambda^m t^m}{(1-\lambda)(m)!} + \frac{\lambda^m t^m}{(1-\lambda)^2 m!} \right] \\
&= \frac{(t-d)e^{-t}}{\lambda^d t} \left[\frac{(\lambda t)e^{\lambda t}}{(1-\lambda)} + \frac{e^{\lambda t}}{(1-\lambda)^2} \right] \\
&= \frac{(t-d)e^{-(1-\lambda)t}}{(1-\lambda)\lambda^{d-1}} + \frac{(t-d)e^{-(1-\lambda)t}}{(1-\lambda)^2 \lambda^d t} && \text{(A 16)}
\end{aligned}
$$

The term $T_{2A(b)}$ is given by

$$
\begin{aligned}
T_{2A(b)} &= \frac{(t-d)e^{-t}}{\lambda^d t}\sum_{k=0}^{d-1}(k+1)\lambda^k\sum_{m=0}^{i+d}\frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}}{\lambda^d t}\sum_{m=0}^{d-1}\sum_{k=m}^{d-1}(k+1)\lambda^k\frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}}{\lambda^d t}\sum_{m=0}^{d-1}\frac{\left[m(1-\backslash)\lambda^m+\lambda^m-d(1-\lambda)\lambda^d-\lambda^d\right]}{(1-\lambda)^2}\frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}}{\lambda^d t}\left[\sum_{m=0}^{d-1}\frac{m\backslash^m t^m}{(1-\lambda)m!}+\sum_{n=0}^{d-1}\frac{t^m\lambda^m}{(1-\lambda)^2}\right. \\
&\quad \left.-\frac{d\lambda^d}{(1-\lambda)}\sum_{m=0}^{d-1}\frac{(t)^m}{m!}-\frac{\lambda^d}{(1-\lambda)^2}\sum_{m=0}^{d-1}\frac{t^m}{m!}\right] \\
&= \frac{(t-d)e^{-t}}{\lambda^d t}\left[\frac{\lambda t}{(1-\lambda)}\sum_{m=0}^{d-2}\frac{(\lambda t)^m}{m!}+\frac{1}{(1-\lambda)^2}\sum_{m=0}^{d-2}\frac{(\lambda t)^m}{m!}\right. \\
&\quad \left.-\frac{d\lambda^d}{(1-\lambda)}\sum_{m=0}^{d-1}\frac{(t)^m}{m!}-\frac{\lambda^d}{(1-\lambda)^2}\sum_{m=0}^{d-1}\frac{t^m}{m!}\right] \\
&= \frac{(t-d)e^{-t}}{\lambda^d t}\left[\frac{\lambda t}{(1-\lambda)}\phi(\lambda t,d-2)+\frac{1}{(1-\lambda)^2}\phi(\lambda t,d-1)\right. \\
&\quad \left.-\frac{d\lambda^d}{(1-\lambda)}\phi(t,d-1)-\frac{\lambda^d}{(1-\lambda)^2}\phi(t\ \ d-1)\right] \\
&= \frac{(t-d)e^{-t}}{(1-\lambda)\lambda^{d-1}}\phi(\lambda t\ \ d-2)+\frac{(t-d)e^{-t}}{(1-\lambda)^2\lambda^d t}\phi(\lambda t\ \ d-1) \\
&\quad -\frac{(t-d)de^{-t}}{(1-\lambda)t}\phi(t\ \ d-1)-\frac{(t-d)e^{-t}}{(1-\lambda)^2 t}\phi(t,d-1) \qquad (A\ 17)
\end{aligned}
$$

The term T_{2B} can be evaluated as given below

$$
\begin{aligned}
T_{2B} &= \frac{(t-d)e^{-t}(d+1)}{\lambda^d t}\sum_{i=0}^{\infty}\lambda^{\ +d}\sum_{m=0}^{i+d}\frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}(d+1)}{\lambda^d t}\sum_{k=d}^{\infty}\lambda^k\sum_{m=0}^{k}\frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}(d+1)}{\lambda^d t}\sum_{k=0}^{\infty}\lambda^k\sum_{m=0}^{k}\frac{t^m}{m!} \\
&\quad -\frac{(t-d)e^{-t}(d+1)}{\lambda^d t}\sum_{k=0}^{d-1}\lambda^k\sum_{m=0}^{k}\frac{t^m}{m!} \\
&= T_{2B(a)}-T_{2B(b)} \qquad (A\ 18)
\end{aligned}
$$

$T_{2B(a)}$ and $T_{2B(b)}$ are evaluated below

$$
T_{2B(a)} = \frac{(t-d)e^{-t}(d+1)}{\lambda^d t}\sum_{k=0}^{\infty}\lambda^k\sum_{m=0}^{k}\frac{t^m}{m!}
$$

$$\begin{aligned}
&= \frac{(t-d)e^{-t}(d+1)}{\lambda^d t}\sum_{m=0}^{\infty}\sum_{k=n}^{\infty}\lambda^k\frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}(d+1)}{\lambda^d t}\sum_{m=0}^{k}\frac{\lambda^m}{(1-\lambda)}\frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}(d+1)}{(1-\lambda)\lambda^d t}\sum_{m=0}^{k}\frac{(\lambda t)^m}{m!} \\
&= \frac{(t-d)(d+1)}{(1-\lambda)\lambda^d t}e^{-(1-\lambda)t} \qquad \text{(A 19)}
\end{aligned}$$

$$\begin{aligned}
T_{2B(b)} &= \frac{(t-d)e^{-t}(d+1)}{\lambda^d t}\sum_{k=0}^{d-1}\lambda^k\sum_{m=0}^{k}\frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}(d+1)}{\lambda^d t}\sum_{m=0}^{d-1}\sum_{k=n}^{d-1}\lambda^k\frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}(d+1)}{\lambda^d t}\sum_{m=0}^{d-1}\frac{(\lambda^m-\lambda^d)}{(1-\lambda)}\frac{t^m}{m!} \\
&= \frac{(t-d)e^{-t}(d+1)}{(1-\lambda)\lambda^d t}\phi(\lambda t, d-1) - \frac{(t-d)e^{-t}(d+1)}{(1-\lambda)t}\phi(t, d-1) \qquad \text{(A 20)}
\end{aligned}$$

The term T_3 can be evaluated as given below

$$\begin{aligned}
T_3 &= \sum_{i=0}^{\infty}\frac{i\lambda^i}{t}e^{-t}\frac{t^{i+d+1}}{(i+d)!} \\
&= \frac{e^{-t}}{\lambda^d}\left[\sum_{i=0}^{\infty}(i+d)\lambda^{i+d}\frac{t^{i+d}}{(i+d)!} - \sum_{i=0}^{\infty}d\lambda^{i+d}\frac{t^{i+d}}{(i+d)!}\right] \\
&= \frac{e^{-t}}{\lambda^d}\left[\sum_{i=0}^{\infty}\frac{(\lambda t)^{\ +d}}{(i+d-1)!} - \sum_{i=0}^{\infty}\frac{d(\lambda t)^{i+d}}{(i+d)!}\right] \\
&= \frac{te^{-t}}{\lambda^{d-1}}\left[e^{\lambda t} - \sum_{i=0}^{d-2}\frac{(\lambda t)^i}{i!}\right] - \frac{de^{-t}}{\lambda^d}\left[e^{\lambda t} - \sum_{i=0}^{d-1}\frac{(\lambda t)^i}{i!}\right] \\
&= \frac{te^{-(1-\lambda)t}}{\lambda^{d-1}} - \frac{te^{-t}}{\lambda^{d-1}}\phi(\lambda t\ \ d-2) - \frac{de^{-(1-\lambda)t}}{\lambda^d} + \frac{de^{-t}}{\lambda^d}\phi(\lambda t, d-1) \qquad \text{(A 21)}
\end{aligned}$$

Then $E[Q_t|D_t = d]$ is given by

$$E[Q_t|D_t = d] \;=\; T_1 - T_{2A(a)} + T_{2A(b)} + T_{2B(a)} - T_{2B(b)} + T_3 \qquad \text{(A 22)}$$

References

[BeS92] Bertsimas,D J and Servi L D , Deducing Queueing from Transactional Data The Queue Inference Engine Revisited *Operations Research* vol 40 supp no 2, pp S217 S228, 1992

[BoC91] Boxma,O J and Cohen J W , 'The M/G/1 Queue with Permanent Customers *IEEE Journal on Selected Areas in Communications* vol 9 no 2 pp 179-184 1991

[Box85] Boxma,O J 'A Queueing Model for Finite and Infinite Source Interaction *Center for Mathematics and Computer Science* Report OS R8511 1985

[COHEN] Cohen J W '*The Single Server Queue* North Holland Publishing Co 1969

[DaS92] Daley,D J and Servi L D , Exploiting Markov Chains to Infer Queue Length from Transactional Data," *Journal of Applied Probability* , vol 29 pp 713 732 1992

[DoW87] Doshi B and Wong,W S 'Exact Solution of a Simple Finite Infinite Source Interaction Model,' *Queueing Systems* vol 2 pp 67 82, 1987

[FELL] Feller W "*An Introduction to Probability Theory and Its Applications Vol 1*, Wiley Eastern Ltd 3rd ed , 1968

[Lar90] Larson,R C , "The Queue Inference Engine Deducing Queue Statistics From Transactional Data,' *Management Science* , vol 36 no 5, pp 586 601, 1990

[KLECN] Kleinrock L , '*Communication Nets Stochastic Message Flow and Delay*, McGraw-Hill, 1964

[KLEIN] Kleınrock L , *Queueıng Systems Volume 1 Theory*, John Wıley & Sons 1975

[Kuma92] Kumar,A , On the Average Idle Tıme and Queue Length Estımates ın an M/M/1 Queue" *Operatıons Research Letters*, vol 12, pp 153 157 1992

[MaM96] Manjunath,D and Molle,M L , Passıve Estımatıon Algorıthms for Queueıng Delays ın LAN s and Other Pollıng Systems *Proc IEEE Infocom 96* vol 1 pp 240 247 1996